FPredict

Flight Delay Prediction with Factor Interpretation

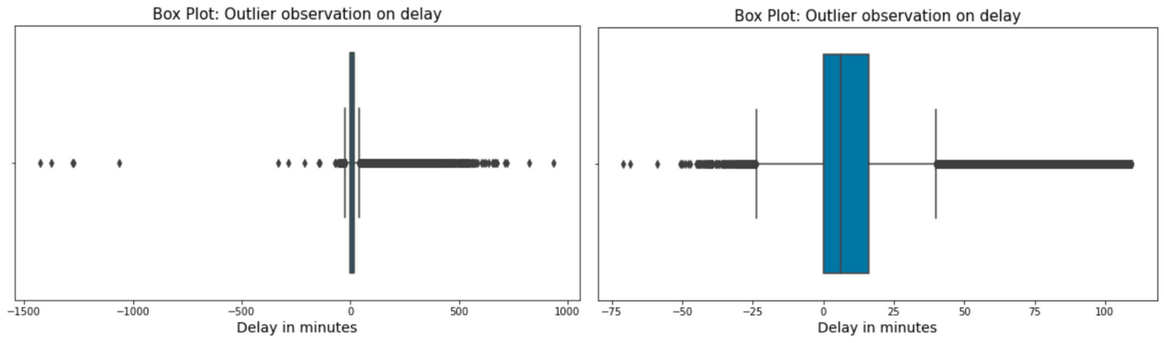
TEAM 77

Gauri Kapse, Shubham Biswas, Kartik Tripathi, Alvin Handoko, Rudra Gupta, Saurabh Tahiliani

Summary

Most flight prediction models focus on either predicting the **extent of a flight's delay,** or providing insights into **factors affecting the extent of a delay**. These also work on a near-real-time basis (24-48 hours). There is currently a lack of a single tool that can **do both functions over a slightly longer horizon.** We developed a **scalable factor-based model** to predict flight delays and identify the most influential factors contributing to the delay. Travellers can use this for flight recommendations and airlines can use it to identify areas for operational optimization.

Data Cleaning and Feature Engineering



Flight delay distribution before (left) and after (right) outlier removal. Negative delays indicate early departures.

Data:

- DXB Flight arrivals and departures data (from *Dubai Pulse, ~50 mb, ~150k rows*)
- Weather data (from *US National Centers for Environmental Information*, ~80 kb, 365 rows per annum)
- 2.5 years (Dec-20 to Mar-23)

Cleaning:

Removed outliers (>3 std.dev from mean), correlated features, duplicated features and imputed missing data

Modelling and Recommendations

Data imbalanced: **most delays under 15 minutes**. Delays bucketed for balancing:

- Delay < 15 minutes = **On Time** (including early arrivals)
- 15 minutes < Delay < 45 minutes = **Delay under 1 hour**
- **Delay over 1 hour** (remaining data)

CatBoost Classifier Model used: automates the conversion of categorical features into numerical features, can train on GPU

Performance metrics of the arrival delay classification model

Class	Precision	Recall	F-1 Score
No delay (9536)	0.82	0.97	0.89
Delay Under 1 Hour (2051)	0.50	0.14	0.21
Delay Over 1 Hour (249)	0.11	0.00	0.01

Recommendations: filter on-time flights from/to the user's chosen location, within a 10 day interval centered on the user's chosen date.

Interpretability

	columns	importance
0	Flight Number	15.067829
1	Destination City	12.250108
2	Parking Position	9.395652
3	Aircraft Type	8.689288
4	Airline	7.521113
5	Mean Temperature	6.775979
6	Dewpoint	5.143626
7	Windspeed	5.065182
8	Day Of Month	4.961617
9	Month	4.456895

Top 10 Features from *Departures* dataset with *CatBoost* model

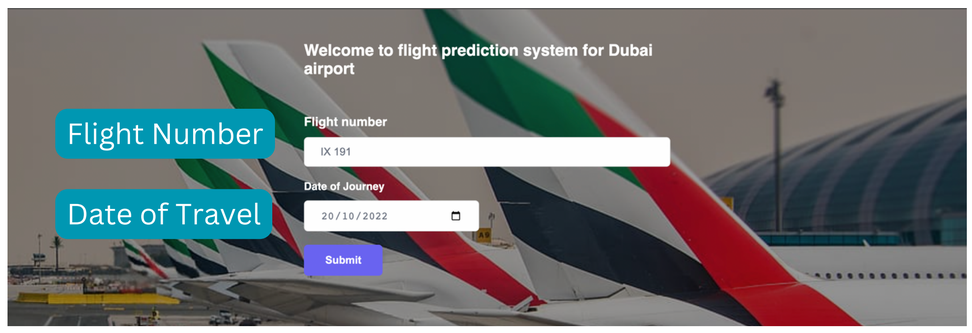
Global interpretability: used feature importance attribute of CatBoost.

Local Interpretability: **SHAP algorithm** (SHapley Additive exPlanations)

App Infrastructure

D3.js for visualization; **Flask** as the backend server.

- **Landing Page**



- **Result Page**

Future Improvements

- CatBoost hyper parameter tuning
- Train model with extended data timeframe
- Real time prediction instead of pre-computed batch prediction
- Obtain forecasted info for a*ircraft parking position* and w*eather*